# Supplementary Appendix
# Landau, Carter & Stojanov et al:

# "Evolution and impact of subclonal mutations in chronic lymphocytic leukemia"

## A. Supplemental Figure Legends

**Figure S1. A. Significantly mutated genes in 160 CLL samples, *related to Figure 1*.** A-S**.** Type (missense, splice-site, nonsense) and location of mutations in the significantly mutated genes discovered among the 160 CLL samples (top) compared to previously reported mutations in literature or in the COSMIC database (v76) (bottom). Dashed boxes in A, C, D, J, O and P indicate mutations localizing to a discrete gene territory. Please refer to previous publication for mutation information for *FBXW7*(Wang et al., 2011) **B.** Mutation sites in 14 significantly mutated genes are localized to conserved regions of genes. Where available, alignments of gene sequences around each mutation are shown for human, mouse, zebrafish, *C.elegans and S.pombe* genes (USCS Genomic Bioinformatics: http://genome.ucsc.edu).

**Figure S2. Whole exome sequencing allelic fraction estimates are consistent with deep sequencing and RNA sequencing measurements, *related to Figure 2*. A.** Comparison of ploidy estimates by ABSOLUTE with flow analyses for DNA content of 7 CLL samples and one normal B cell control (not analyzed by ABSOLUTE). Vertical lines indicate 95% confidence intervals of ploidy measurements by FACS. **B.** Comparison of measurements of allelic fraction of 256 gene mutations detected by WES compared to detection using Fluidigm-based amplification following by deep sequencing (average 4200x coverage) using a MiSeq instrument. Significantly different estimates were assigned open circles. **C.** Comparison of allelic fraction measured for 74 validated sites from 16 CLL samples by WES or RNA sequencing. **D**. Comparison of mutational spectrum between subclonal and clonal sSNVs (detected in 149 CLLs). Rates were calculated as the fraction of the total number of sSNVs in the set with a particular mutation variant.

**Figure S3. Co-occurrence of mutations, *related to Figure 3*.** The commonly occurring mutations, sorted in the order of decreasing frequency of affected. The top panel - the total number of mutations (red) and the number of subclonal mutations (blue) per sample. Bottom panel - co-occurring CLL driver events (y-axis) are marked per individual CLL sample (x-axis). Color spectrum (light yellow to black) corresponds to CCF; white boxes - no driver mutation identified; grey - mutations whose CCF was not estimated (i.e. mutations involving the X chromosome and indels other than in NOTCH1, currently not evaluated with ABSOLUTE).

**Figure S4. Characterization of CLL clonal evolution through analysis of subclonal mutations at two timepoints in 18 patients, *related to Figure 4*. A-B.** Unclustered results for 18 longitudinally studied CLLs, comparing CCF at two timepoints, Red color denotes a mutation with an increase in CCF greater than 0.2 (with probability >0.5). Six CLLs with no interval treatment (**A**) and 12 CLLs with intervening treatment (**B**) were classified as non-evolvers or evolvers, based on the presence of mutations with a statistically significant increase in CCF. **C.** Deep sequencing validation of 6 of the 18 CLLs. For each set of samples, allelic frequency (AF) by WES (red) (with 95% CI by binofit shown by cross bars) is shown on the left and AF by deep sequencing (blue) (with 95% CI by binofit shown by cross bars) is shown on the right. Deep sequencing was performed to an average coverage of 4200x. **D.** RNA pyrosequencing demonstrates a change in mRNA transcript levels that are consistent with changes in DNA allelic

frequencies. **E.** Genetic changes correlate with transcript level of pre-defined gene sets expected to be altered as a result of the genetic lesion. These include change in expression level in the nonsense-mediated mRNA decay (NMD) pathway gene set, expected to be increased in association with splicing abnormalities such as *SF3B1* mutations (**Table S10B**). In addition, changes in expression level of the NRASQ61 gene set (**Table S10A**) accompany the shift in allelic frequency for the *NRAS* mutations.

**Figure S5. The presence of a subclonal driver is associated with shorter FFS_Sample when added to known clinical high risk indicators (*related to Figure 6*).** FFS_Sample plots of the patient groups based on presence or absence of a subclonal driver ('+/- SC driver') and their **(A)** *IGHV* mutation status; **(B)** exposure to prior therapy; **(C)** presence or absence of del(11q) and **(D)** presence or absence of del(17p).

## B. Supplemental Tables

**Table S1.** Summary of the clinical characteristics of 160 patients whose leukemia samples underwent whole-exome sequencing and of the associations between patient characteristics and the number of clonal or subclonal mutations detected per sample for the 149 patients for whom ABSOLUTE output was available.

| | WES N (%) | N (%) | Number of Clonal Muts Median (Range) | *P*-value‡ | Number of Subclonal Muts Median (Range) | *P*-value‡ |
|---|---|---|---|---|---|---|
| N | 160 | 149 | | | | |
| Age (median =54) | | | | | | |
| <54 yrs. | 87 (64) | 72 (46) | 7.5 (0, 21) | <0.001 | 7 (0, 30) | 0.63 |
| ≥54 yrs. | 73 (46) | 77 (54) | 12 (1, 30) | | 8 (0, 25) | |
| Sex | | | | | | |
| Female | 61 (38) | 59 (40) | 10 (0, 27) | 0.60 | 7 (0, 25) | 0.46 |
| Male | 99 (62) | 90 (60) | 9.5 (1, 30) | | 8 (0, 30) | |
| Rai Stage at Sample | | | | | | |
| 0-1 | 117 (73) | 110 (74) | 10 (0, 30) | 0.081 | 7 (0, 25) | 0.070 |
| 2-4 | 37 (23) | 34 (23) | 9 (1, 19) | | 8.5 (1, 30) | |
| Unknown | 6 (4) | 5 (3) | | | | |
| Treatment Status at time of Sample | | | | | | |
| Chemotherapy naïve | 127 (79) | 120 (81) | 9.5 (1, 30) | 0.63 | 7 (0, 25) | 0.015 |
| Prior Treatment | 33 (21) | 29 (19) | 10 (0, 27) | | 10 (1, 30) | |
| Number of Prior Therapies at Sample | | | | | | |
| 0 | 127 (79) | 120 (81) | 9.5 (1, 30) | 0.88 | 7 (0, 25) | 0.011 |
| 1-3 | 27 (17) | 24 (16) | 10.5 (0, 19) | | 9 (1, 29) | |
| 4-6 | 6 (4) | 5 (3) | 7 (4, 27) | | 21 (5, 30) | |
| *IGHV* status | | | | | | |
| Mutated | 84 (53) | 82 (55) | 11 (1, 30) | 0.054 | 7 (0, 25) | 0.61 |
| Unmutated | 51 (32) | 47 (32) | 9 (0, 19) | | 8 (0, 29) | |
| Unknown | 25 (15) | 20 (13) | | | | |
| ZAP-70 expression | | | | | | |
| Negative | 53 (33) | 73 (49) | 10 (0, 30) | 0.78 | 7 (0, 29) | 0.55 |
| Positive | 76 (48) | 47 (32) | 10 (1, 27) | | 8 (0, 30) | |
| Unknown | 31 (19) | 29 (19) | | | | |
| FISH (Dohner Classification)† | | | | | | |
| del(17p) (worst) | 20 (13) | 19 (13) | 13 (5, 27) | 0.002 | 7 (1, 29) | 0.62 |
| del(11q) | 24 (15) | 21 (14) | 9 (1, 12) | | 9 (0, 30) | |
| Trisomy 12 | 14 (9) | 16 (11) | 7 (2, 16) | | 7 (1, 19) | |
| Normal karyotype | 23 (14) | 22 (15) | 8 (0, 27) | | 7 (1, 16) | |
| del(13q) (best) | 58 (36) | 56 (38) | 11 (1, 30) | | 7 (0, 25) | |
| Unknown | 20 (13) | 15 (10) | | | | |

†Data known for 134 patients;15 unknown. ǂTest excludes unknowns

**Table S6.** Clinical characteristics of 18 patients for whom longitudinal samples were studied.

| CLL IDs | Age | Therapy | | *IGHV* mutation status | *ZAP70* status | FISH Cytogenetics | Years between samples |
|---|---|---|---|---|---|---|---|
| | | At timepoint 1 | Between timepoints 1 & 2 | | | | |
| CLL018 | 71 | None | None | Y | - | del(13q) | 2.4 |
| CLL026 | 50 | None | None | Y | - | del(13q) | 4.5 |
| CLL020 | 54 | None | None | Y | + | del(13q) | 2.5 |
| CLL019 | 52 | None | None | Y | - | del(13q) | 3.2 |
| CLL030 | 54 | None | None | Y | + | del(13q) | 3.5 |
| CLL082 | 77 | None | None | N | + | del(13q,17p), tri12 | 3.1 |
| CLL011 | 41 | None | FCR | N | + | del(13q) | 5 |
| CLL088 | 60 | None | FCR, Alem+R | N | - | tri12 | 4.5 |
| CLL116 | 36 | None | FCR | N | + | tri12 | 4.3 |
| CLL169 | 69 | None | FR | Y | + | del(13q) | 4.7 |
| CLL167 | 56 | None | FR | Y | - | del(13q),tri12 | 2.7 |
| CLL016 | 59 | None | FR | N | + | del(13q) | 3.4 |
| CLL001 | 58 | None | FR | N | + | del(11q, 13q) | 3.5 |
| CLL065 | 75 | FCR | Alem+R,R | N | + | del(13q),del(17p) | 3.1 |
| CLL006 | 67 | FC, Chloram | Alem+R, FR, exp. | N | - | del(13q),del(11q) | 4.6 |
| CLL014 | 65 | R | FR | Y | - | del(13q) | 2.9 |
| CLL066 | 70 | FR, Chloram | R-CVP | Y | - | del(13q) | 3.5 |
| CLL040 | 60 | FCR | FCR, Alem+R | N | + | del(13q),del(11q) | 3 |

Abbreviations: Y- Yes, N- No, Mut.- Mutated, FISH-Fluorescence In Situ Hybridization, F-Fludarabine, C- Cyclophosphamide, R-Rituximab, V-Vincristine, Chloram- Chlorambucil, Alem – Alemtuzumab; Rev – Revlimid; exp - experimental

**Table S8.** The 149 CLL samples could be divided into two groups, based on presence or absence of subclonal driver. **A**. Patient characteristics; **B.** Driver events that participated in the analysis of subclonal drivers as a predictor of clinical outcome (*related to **Figure 6***)

**A**. **Patient characteristics at diagnosis**

| | **Total** | **Subclonal driver not-detected** | **Subclonal driver detected** | ***P*-value** |
|---|---|---|---|---|
| | **N (%)** | **N (%)** | **N (%)** | |
| **N** | 149 | 81 (54) | 68 (46) | |
| **Age (years) at Diagnosis, median (range)** | 54 (34, 77) | 55 (34, 77) | 53 (36, 76) | 0.52 |
| **Age ≥54 yrs.** | 77 (52) | 45 (56) | 32 (47) | 0.33 |
| **Sex** | | | | |
| **Female** | 59 (40) | 25 (30) | 34 (51) | 0.02 |
| **Male** | 90 (60) | 56 (69) | 34 (50) | |
| **Rai Stage at Sample** | | | | |
| **0-1** | 110 (74) | 63 (78) | 47 (69) | 0.12 |
| **2-4** | 34 (23) | 14 (17) | 20 (29) | |
| **Unknown** | 5 (3) | 4 (5) | 1 (1) | |
| **Treatment Status at time of Sample** | | | | |
| **Chemotherapy naïve** | 120 (81) | 73 (90) | 47 (69) | 0.002 |
| **Prior Treatment** | 29 (19) | 8 (10) | 21 (31) | |
| **Number of Prior Therapies at Sample** | | | | |
| **0** | 120 (81) | 73 (90) | 47 (69) | 0.003 |
| **1-3** | 24 (16) | 7 (9) | 17 (25) | |
| **4-6** | 5 (3) | 1 (1) | 4 (6) | |

**B. Driver genetic events used in the analysis of subclonal drivers as a predictor of clinical outcome.** For each driver, the number of CLL samples out of the cohort of 149 harboring a subclonal specified event is provided. Driver events were defined as being included within the Cancer Gene Census(Futreal et al., 2004) ('CGC') and affecting a highly conserved location; or as within the list of significantly occurring events in CLL (**Figure 1A**). For CGC and CLL_Drivers, 1-yes; 0-no.

| Driver ID | Number of samples harboring a subclonal driver | CGC | CLL_Driver |
|---|---|---|---|
| *ATM* | 4 | 1 | 1 |
| *CHD2* | 4 | 0 | 1 |
| *EGR2* | 1 | 0 | 1 |
| *FBXW7* | 3 | 1 | 1 |
| *ITPKB* | 1 | 0 | 1 |
| *KRAS* | 3 | 1 | 1 |
| *NRAS* | 3 | 1 | 1 |
| *POT1* | 2 | 0 | 1 |
| *RIPK1* | 1 | 0 | 1 |
| *SF3B1* | 10 | 1 | 1 |
| *TP53* | 9 | 1 | 1 |
| *XPO1* | 1 | 1 | 1 |
| *DICER1* | 1 | 1 | 0 |
| *IDH1* | 2 | 1 | 0 |
| *MLL2* | 1 | 1 | 0 |
| *NOTCH1* | 4 | 1 | 1 |
| *NSD1* | 1 | 1 | 0 |
| *PBRM1* | 1 | 1 | 0 |
| *PTPN11* | 1 | 1 | 0 |
| *SDHB* | 1 | 1 | 0 |
| *TSC2* | 1 | 1 | 0 |
| del(11q) | 12 | 0 | 1 |
| del(13q) (het) | 18 | 0 | 1 |
| del(13q) (hom) | 7 | 0 | 1 |
| del(17p) | 8 | 0 | 1 |
| del(8p) | 5 | 0 | 1 |
| trisomy 12 | 1 | 0 | 1 |

**Table S9. Statistical modeling for the effect of the presence of clonal and subclonal drivers on FFS_Sample and FFS_Rx**.

**A.** Kaplan Meier analysis for the effect of the presence of a subclonal or a clonal driver on the following 2 outcome measures: (i) FFS_Sample: time from WES sampling to the first treatment after sampling or death, and (ii) FFS-Rx: time from the first treatment after sampling to the second treatment after sampling or death. FFS_Sample was analyzed in 149 patients, while FFS-Rx was analyzed in 67 patients who were treated after sampling. Kaplan Meier analysis of FFS-Sample restricted to 132/149 CLL samples with at least 1 driver event (irrespective of cancer cell fraction) detected and of FFS-Rx restricted to 62 CLLs patients treated after sampling with at least 1 driver event is also shown.

| | N | Median FFS_Sample Months (95% CI)† | *P*-value | N | Median FFS_Rx Months (95% CI)† | *P*-value |
|---|---|---|---|---|---|---|
| N | 149 | | | 67 | | |
| Clonal Driver | | | | | | |
| Absent | 36 | NR (NR, NR) | 0.026 | 12 | NR (13, NR) | 0.38 |
| Present | 113 | 27 (16, NR) | | 55 | 43 (33, 48) | |
| Subclonal Driver | | | | | | |
| Absent | 81 | NR (49, 74) | <0.001 | 28 | NR (44, NR) | 0.006 |
| Present | 68 | 15 (6, 27) | | 39 | 33 (14, 45) | |
| Including only Patients with at least one driver identified | 132 | | | 62 | | |
| Clonal Driver | | | | | | |
| Absent | 19 | NR (15, NR) | 0.20 | 7 | 18 (13, NR) | 0.78 |
| Present | 113 | 27 (16, NR) | | 55 | 43 (33, 48) | |
| Subclonal Driver | | | | | | |
| Absent | 64 | NR (29, NR) | <0.001 | 23 | 48 (35, NR) | 0.021 |
| Present | 68 | 15 (6, 27) | | 39 | 33 (14, 45) | |

†NR=Median not reached

**B.** Adjusted Cox Modeling for FFS_Rx for known high-risk prognostic features in CLL restricted to the 62 patients who had at least one driver identified

| | HR (95% CI) | p-value |
|---|---|---|
| FFS_Rx | | |
| **Subclonal** driver present vs. absent | 3.34 (1.31, 8.52) | 0.012 |
| IGHV Unmutated vs. Mutated | 2.02 (0.69, 5.91) | 0.20 |
| Missing vs. Mutated | 1.37 (0.25, 7.51) | 0.72 |
| del(11q) present vs. absent | 1.40 (0.57, 3.45) | 0.47 |
| del(17p) present vs. absent | 5.03 (1.77, 14.30) | 0.002 |
| Prior therapy vs. None | 1.65 (0.72, 3.81) | 0.24 |

**C.** Adjusted Cox Modeling for FFS_Sample for known high-risk prognostic features in CLL

| | HR (95% CI) | p-value |
|---|---|---|
| FFS_Sample | | |
| **Subclonal** driver present vs. absent | 1.66 (0.98, 2.79) | 0.058 |
| IGHV Unmutated vs. Mutated | 2.87 (1.54, 5.36) | <0.001 |
| Missing vs. Mutated | 1.02 (0.39, 2.63) | 0.98 |
| del(11q) present vs. absent | 2.03 (1.11, 3.70) | 0.021 |
| del(17p) present vs. absent | 1.22 (0.65, 2.28) | 0.54 |
| Prior therapy vs. None | 3.42 (1.78, 6.57) | <0.001 |
| **Clonal** Driver vs. None | 1.25 (0.65, 2.41) | 0.50 |
| IGHV Unmutated vs. Mutated | 2.99 (1.60 5.58) | <0.001 |
| Missing vs. Mutated | 1.01 (0.40, 2.55) | 0.99 |
| del(11q) vs. None | 1.95 (1.07, 3.53) | 0.029 |
| del(17p) vs. None | 1.13 (0.61, 2.08) | 0.70 |
| Prior Trt vs. None | 3.98 (2.13, 7.43) | <0.001 |

**D.** Adjusted and unadjusted hazard ratios (HR) for the effect of the number of subclonal or clonal drivers on FFS_Rx

| | Unadjusted HR (95% CI) | *P*-value | Adjusted† HR (95% CI) | *P*-value |
|---|---|---|---|---|
| FFS_Rx | | | | |
| No. of Subclonal Driver Events | | | | |
| 1-2 vs. none | 2.73 (1.11, 6.70) | 0.029 | 3.46 (1.28, 9.40) | 0.015 |
| 3-5 vs. none | 7.31 (2.28, 23.40) | <0.001 | 4.02 (1.08, 14.95) | 0.038 |
| No. of Clonal Driver Events | | | | |
| 1-2 vs. none | 1.73 (0.52, 5.76) | 0.37 | 1.29 (0.37, 4.52) | 0.69 |
| 3-5 vs. none | 1.35 (0.22, 8.11) | 0.74 | 0.45 (0.07, 3.10) | 0.42 |

†Adjusted for *IGHV* mutation status, del(17p), del(11q), prior treatment at time of sample

**E.** Cox Modeling for FFS_Rx adjusting for high-risk mutations (*TP53*, *ATM* and *SF3B1*)

| | HR (95% CI) | *P*-value |
|---|---|---|
| FFS_Rx | | |
| Subclonal Driver vs. None | 2.76 (1.08, 7.01) | 0.033 |
| *SF3B1* present vs. absent | 0.83 (0.27, 2.58) | 0.74 |
| *ATM* present vs. absent | 1.94 (0.54, 6.98) | 0.31 |
| *TP53* present vs. absent | 8.75 (3.14, 24.39) | <0.001 |

**Table S10A.** Geneset and expression values for 44 genes specific for *NRAS* Q61 mutations. These genes were compiled by Eskandarpour, M., et al (Table S1 (Eskandarpour et al., 2009)) and reflect genes downregulated by more than -1.41 fold after RNAi-mediated silencing of the *NRAS* Q61 mutant in 224 cell lines. Tumor cells from CLL088 harbored an *NRAS* Q61 mutation, and displayed is the difference in gene expression between timepoint 1 vs. timepoint 2 of CLL088. Average and standard deviation of gene expression difference between timepoints 1 and 2 is given for all other longitudinal sample pairs.

| Gene symbol | CLL088 | All other longitudinal samples (17 pairs) | |
|---|---|---|---|
| | | Average | Standard deviation |
| *NPPC* | -0.04578799 | -0.007210208 | 0.514616415 |
| *COL13A1* | 0.134673459 | -0.04659692 | 0.254756387 |
| *CDC42EP3* | 0.959816072 | 0.082498719 | 0.396245179 |
| *ENC1* | -0.131627188 | -0.008812696 | 0.560586955 |
| *SET* | 0.243856478 | 0.017019055 | 0.163422427 |
| *TMEM2* | -0.633168048 | 0.056473858 | 0.486765405 |
| *CTGF* | 0.446446876 | 0.062380273 | 0.350426395 |
| *PHLDA1* | 0.0856608 | -0.11617691 | 0.242079319 |
| *PVRL3* | 0.051635241 | -0.071360125 | 0.309077828 |
| *NEDD4* | 0.372647977 | -0.065753321 | 0.293381482 |
| *HMGA1* | 0.721592229 | 0.014810233 | 0.294059124 |
| *IGFBP3* | 0.285836175 | 0.113474904 | 0.221068564 |
| *SPRY4* | -0.427121219 | 0.05373157 | 0.195213896 |
| *DLD* | 0.658330384 | 0.0623254 | 0.339485653 |
| *HCCS* | 0.29885353 | -0.07203308 | 0.356171393 |
| *DPYSL2* | 0.821204985 | -0.053448441 | 0.973626504 |
| *EPHA2* | -0.310446211 | -0.061035231 | 0.350962757 |
| *THBS1* | 0.614908738 | 0.237633954 | 0.398663855 |
| *LSM7* | 0.243760925 | 0.059988504 | 0.3344074 |
| *NIP7* | 1.252496702 | 0.049585749 | 0.502407045 |
| *MGLL* | 2.768195308 | 0.015144031 | 0.390774595 |
| *PTX3* | 1.171309917 | 0.313612997 | 0.683123534 |
| *KCNN4* | -0.47768552 | -0.100262781 | 0.391350388 |
| *LPXN* | 0.300270992 | -0.035757494 | 0.350558014 |
| *CDC42EP3.1* | 0.959816072 | 0.082498719 | 0.396245179 |
| *DUSP6* | -0.782740031 | 0.577139524 | 1.037845647 |
| *ETF1* | 0.531177079 | 0.026704345 | 0.446412676 |
| *LEF1* | -0.112434288 | -0.031090606 | 0.737631178 |
| *RHOBTB3* | 0.088565447 | 0.087647222 | 0.361094271 |
| *CPD* | 0.338056607 | 0.242027513 | 0.314157079 |
| *MKRN2* | 0.048507309 | -0.027646317 | 0.234771743 |
| *PLAUR* | 1.568611765 | 0.428891808 | 0.488058773 |
| *STK17A* | -0.852885509 | 0.127270233 | 0.435271027 |
| *CCDC85B* | 0.125112267 | -0.132016138 | 0.303754024 |

| | | | |
|---|---|---|---|
| *FZD7* | 0.760598355 | -0.135184117 | 0.510611289 |
| *PLAUR.1* | 1.568611765 | 0.428891808 | 0.488058773 |
| *SET.1* | 0.243856478 | 0.017019055 | 0.163422427 |
| *CCNE2* | 0.302043507 | 0.031955004 | 0.262347119 |
| *FNDC3B* | 0.493904857 | 0.216358443 | 0.228289601 |
| *IFRD1* | 0.419114097 | -0.142779223 | 0.483899647 |
| *DDX21* | 0.616773364 | 0.096081169 | 0.357469727 |
| *FTSJ1* | 0.361269841 | -0.012203281 | 0.190033873 |
| *PDLIM5* | 0.567478758 | 0.044486012 | 0.269337671 |
| *RRM2* | 1.566222249 | 0.205188047 | 0.698854981 |
| **Average** | **0.414030014** | **0.059806165** | **0.121855537** |

**Table S10B**. Geneset and expression values for 34 genes involved in RNA export from the nucleus and extended nonsense mRNA degradation (NMD), that have been associated with mutated *SF3B1* (taken from Table S7 of the report by Yoshida et al.(Yoshida et al., 2011)). Tumor cells from CLL040 harbored a spliceosome mutation, *SF3B1;* displayed is the difference in gene expression between timepoint 1 vs. timepoint 2 of CLL040. Average and standard deviation of gene expression difference between timepoints 1 and 2 is given for all other longitudinal sample pairs.

| **Gene symbol** | **CLL040** | **All other longitudinal samples (17 pairs)** | |
|---|---|---|---|
| | | **Average** | **Standard deviation** |
| *SMG5* | 0.011920117 | 0.023259127 | 0.178168824 |
| *DHX34* | 0.025891448 | -0.062902083 | 0.116567172 |
| *UPF1* | 0.21578735 | -0.055402974 | 0.233816043 |
| *SMG1* | 0.570738034 | -0.000372049 | 0.282293591 |
| *UPF3B* | 0.078425721 | -0.047789483 | 0.24001489 |
| *UPF2* | 0.82684527 | 0.058484108 | 0.1991368 |
| *SMG7* | 0.137850743 | -0.038354666 | 0.191155824 |
| *MAGOH* | -0.572093408 | 0.091763238 | 0.536068491 |
| *SMG6* | -0.28288667 | 0.027035851 | 0.147910467 |
| *UPF3A* | 0.426259455 | -0.0991856 | 0.371500187 |
| *CASC3* | 0.178042037 | 0.04766774 | 0.104753942 |
| *RBM8A* | 0.190297511 | 0.028176467 | 0.201891218 |
| *WIBG* | 0.024652918 | -0.065314855 | 0.295338558 |
| *EIF4A3* | -0.00796301 | 0.04062118 | 0.457332467 |
| *SMG5.1* | 0.011920117 | 0.023259127 | 0.178168824 |
| *UPF1.1* | 0.21578735 | -0.055402974 | 0.233816043 |
| *DDX19B* | -0.129049993 | 0.05381727 | 0.26148355 |
| *SMG1.1* | 0.570738034 | -0.000372049 | 0.282293591 |
| *NCBP2* | 0.213822866 | -0.004671586 | 0.257585504 |
| *DDX39* | 0.411741289 | -0.012796232 | 0.283422242 |
| *UPF2.1* | 0.82684527 | 0.058484108 | 0.1991368 |
| *NUDT4* | 0.50454556 | -1.95E-05 | 0.248457474 |

| | | | |
|---|---|---|---|
| *SMG7.1* | 0.137850743 | -0.038354666 | 0.191155824 |
| *RAE1* | 0.067685754 | -0.017983112 | 0.214508052 |
| *SMG6.1* | -0.28288667 | 0.027035851 | 0.147910467 |
| *TSC1* | 0.44211296 | -0.00022797 | 0.276786249 |
| *BAT1* | 0.685210093 | -0.080364706 | 0.270842978 |
| *KHDRBS1* | -0.160256707 | -0.013673409 | 0.193125482 |
| *NUP160* | 0.909274188 | -0.026037476 | 0.392210773 |
| *NXF5* | -0.212514266 | 0.102488019 | 0.250391443 |
| *EIF5A* | -0.14573319 | -0.117388335 | 0.568603045 |
| *DDX25* | 0.527897747 | -0.033354581 | 0.208245641 |
| *NUP133* | 0.006008087 | -0.031510117 | 0.35409961 |
| *NUP107* | 0.127921893 | -0.040181016 | 0.263752523 |
| **Average** | **0.192726137** | **-0.007634335** | **0.047087437** |

## C. Extended Experimental Procedures

**Human samples:** Heparinized blood, skin biopsies and saliva were obtained from patients enrolled on clinical research protocols at the Dana-Farber Harvard Cancer Center (DFHCC) approved by the DFHCC Human Subjects Protection Committee. The diagnosis of CLL according to WHO criteria was confirmed in all cases by flow cytometry, or by lymph node or bone marrow biopsy. Peripheral blood mononuclear cells (PBMC) from normal donors and patients were isolated by Ficoll/Hypaque density gradient centrifugation. Mononuclear cells were used fresh or cryopreserved with FBS 10% DMSO and stored in vapour-phase liquid nitrogen until the time of analysis. Primary skin fibroblast lines were generated from skin punch biopsies as previously described(Wang et al., 2011). The patients included in the cohort represent the broad clinical spectrum of CLL (**Table S1**).

**Established CLL prognostic factor analysis:** *Immunoglobulin heavy-chain variable* (*IGHV*) homology (unmutated was defined as greater than or equal to 98% homology to the closest germline match) and *ZAP-70* expression (high risk defined as >20% positive) were determined(Rassenti et al., 2008). Cytogenetics were evaluated by FISH for the most common CLL abnormalities (del(13q), trisomy 12, del(11q), del(17p), rearrangements of chromosome 14) (all probes from Vysis, Des Plaines, IL, performed at the Brigham and Women's Hospital Cytogenetics Laboratory, Boston MA). Samples were scored positive for a chromosomal aberration based on consensus cytogenetic scoring(Smoley et al., 2010).

**DNA quality control:** We used standard Broad Institute protocols as recently described (Berger et al., 2011; Chapman et al., 2011). Tumor and normal DNA concentration were measured using PicoGreen® dsDNA Quantitation Reagent (Invitrogen, Carlsbad, CA). A minimum DNA concentration of 60 ng/μl was required for sequencing. In select cases where concentration was <60 ng/μl, ethanol precipitation and re-suspension was performed. Gel electrophoresis confirmed that the large majority of DNA was high molecular weight. All Illumina sequencing libraries were created with the native DNA. The identities of all tumor and normal DNA samples (native and WGA product) were confirmed by mass spectrometric fingerprint genotyping of 24 common SNPs (Sequenom, San Diego, CA).

**Whole-exome DNA sequencing:** Informed consent on DFCI IRB-approved protocols for whole exome sequencing of patients' samples was obtained prior to the initiation of sequencing studies. DNA was extracted from blood or marrow-derived lymphocytes (tumor) and saliva, fibroblasts or granulocytes (normal), as previously described (Wang et al., 2011). Libraries for whole exome (WE) sequencing were constructed and sequenced on either an Illumina HiSeq 2000 or Illumina GA-IIX using 76 bp paired-end reads. Details of whole exome library construction have been detailed elsewhere (Fisher et al., 2011). Standard quality control metrics, including error rates, percentage passing filter reads, and total Gb produced, were used to characterize process performance before

downstream analysis. Average exome coverage depth was 132x/146x for tumor/germline. The Illumina pipeline generates data files (BAM files) that contain the reads together with quality parameters. Of the 160 CLL samples reported in the current manuscript, 82 were included in a previous study (Wang et al., 2011). 340 CLL and germline samples were sequenced overall. These include 160 CLL and matched germline DNA samples as well as timepoint 2 samples for 17 of 160 CLLs, and an additional sample pair and germline for a longitudinal sample pair not included in the 160 cohort (CLL020).

**Identification of somatic mutations:** Output from Illumina software was processed by the "Picard" data processing pipeline to yield BAM files containing aligned reads (via MAQ, to the NCBI Human Reference Genome Build hg18) with well-calibrated quality scores (Chapman et al., 2011; DePristo et al., 2011). For 51 of the 160 CLL samples included in the analysis, sequencing was performed on capture libraries generated from whole genome amplified (WGA) samples. For those samples, 100 ng inputs of samples were whole genome amplified with the Qiagen REPLI-g Midi Kit (Valencia, CA).
From the sequencing data, somatic alterations were identified using a set of tools within the "Firehose" pipeline, developed at the Broad Institute (www.broadinstitute.org/cancer/cga). The details of our sequencing data processing have been described elsewhere (Berger et al., 2011; Chapman et al., 2011). Somatic single nucleotide variations (sSNVs) were detected using MuTect [V119, http://www.broadinstitute.org/cancer/cga/mutect, (Cibulskis et al, under review)]; somatic small insertions and deletions (indels) were detected using Indelocator [v61, http://www.broadinstitute.org/cancer/cga/indelocator, (Wang et al., 2011)]. All mutations identified in longitudinal samples were confirmed by manual inspection of the sequencing data (Robinson et al., 2011). An estimated contamination threshold of 5% was used for all samples based on the highest contamination values seen in a formal contamination analysis done with ContEst based on matched SNP arrays (Cibulskis et al., 2011). Ig loci mutations were not included in this analysis. All somatic mutations detected in the 160 CLL samples are listed in **Table S2**. WES data is deposited in dbGaP (phs000435.v1.p1).

**Significance analysis for recurrently mutated genes:** The prioritization of somatic mutations in terms of conferring selective advantage was done with the statistical method MutSig2.0 (Lohr et al., 2012). In short, the algorithm takes an aggregated list of mutations and tries to detect genes that are affected more than expected by chance, as those likely reflect positive selection (i.e driver events). There are two main components to MutSig2.0:

1. The first component attempts to model the background mutation rate for each gene, while taking into account various different factors. Namely, it takes into account the fact that the background mutation rate may vary depending on the base context and base change of the mutation, as well as the fact that the background rate of a gene can also vary across different patients. Given these factors and the background model, it uses convolutions of binomial distributions to calculate a *P* value, which represents the probability that we obtain the

observed configuration of mutations, or a more significant one.

2. The second component of the algorithm focuses on the positional configuration of mutations and their sequence conservation (Lohr et al., 2012). For each gene, the algorithm permutes the mutations preserving their tri-nucleotide context, and for each permutation calculates two metrics: one that measures the degree of clustering into hotspots along the coding length of the gene, and one that measures the average conservation of mutations in the gene. These two null models are then combined into a joint distribution, which is used to calculate a *P* value that reflects the probability by chance that we can obtain by chance the observed mutational degree of clustering and conservation, or a more significant outcome.

The two *P* values that are produced by the two components are then combined using Fisher-Combine (Fisher, 1932) which yields a final *P* value which is used to sort the genes by degree of mutational significance. This is subsequently corrected for multi-hypothesis using the Benjamini Hochberg procedure.

**Genome-wide copy number analysis**: Genome-wide copy number profiles of 111 CLL samples and their patient-matched germline DNA were obtained using the Genome-wide Human SNP Array 6.0 (Affymetrix), according to the manufacturer's protocol (Genetic Analysis Platform, Broad Institute, Cambridge MA). SNP array data were deposited in dbGaP (phs000435.v1.p1). Allele-specific analysis also allowed for the identification of copy neutral LOH events as well as quantification of the homologous copy-ratios (HSCSs) [HAPSEG(Carter, 2011)]. Significant recurrent chromosomal abnormalities were identified using the GISTIC2.0 algorithm ((Mermel et al., 2011),v87). Regions with germline copy number variants were excluded from the analysis.

For CLL samples with no available SNP arrays (38/160), sCNAs were estimated directly from the WES data, based on the ratio of CLL sample read-depth to the average read-depth observed in normal samples for that region. 11/160 samples were excluded from this analysis due to inability to obtain copy number information from the WES data. See **Fig. 2A** for outline of sample processing.

**Validation deep sequencing:** Validation targeted resequencing of 256 selected somatic mutations sSNVs was performed using microfluidic PCR. Target specific primers with Fluidigm-compatible tails were designed to flank sites of interest and produce amplicons of 200 +/-20bp. Molecular barcoded, Illumina-compatible oligonucleotides, containing sequences complementary to the primer tails were added to the Fluidigm Access Array chip (San Francisco, CA) in the same well as the genomic DNA samples (20 - 50 ng of input) such that all amplicons for a given genomic sample shared the same index, and PCR was performed according to the manufacturer's recommendations. Indexed libraries were recovered for each sample in a single collection well on the Fluidigm chip, quantified using picogreen and then normalized for uniformity across libraries. Resulting normalized libraries were loaded on a MiSeq instrument (Illumina) and sequenced using paired end 150bp sequencing reads. 95.2% of called sSNVs were

detected in the validation experiment (**Table S4**). For 91.8% of the mutations, the allelic fraction estimates were concordant (with the discordant events enriched in sites of lower WES coverage).

**RNA sequencing (dUTP Library Construction):** 5μg of total RNA was poly-A selected using oligo-dT beads to extract the desired mRNA. The purified mRNA is treated with DNAse, and cleaned up using SPRI (Solid Phase Reversible Immobilization) beads according to the manufacturers' protocol. Selected Poly-A RNA was then fragmented into ~450 bp fragments in an acetate buffer at high heat. Fragmented RNA was cleaned with SPRI and primed with random hexamers before first strand cDNA synthesis. The first strand was reverse transcribed off the RNA template in the presence of Actinomycin D to prevent hairpinning and purified using SPRI beads. The RNA in the RNA-DNA complex was then digested using RNase H. The second strand was next synthesized with a dNTP mixture in which dTTPs had been replaced with dUTPs. After another SPRI bead purification, the resultant cDNA was processed using Illumina library construction according to manufacturers protocol (end repair, phosphorylation, adenylation, and adaptor ligation with indexed adaptors). SPRI-based size selection was performed to remove adapter dimers present in the newly constructed cDNA library. Libraries were then treated with Uracil-Specific Excision Reagent (USER) to nick the second strand at every incorporated Uracil (dUTP). Subsequently, libraries were enriched with 8 cycles of PCR using the entire volume of sample as template. After enrichment, the library is quantified using pico green, and the fragment size is measured using the Agilent Bioanalyzer according to manufactures protocol. Samples were pooled and sequenced using either 76 or 101bp paired end reads.

**RNASeq data analysis:** RNAseq BAMs were aligned to the hg18 genome using the TopHat suite. Each somatic base substitution detected by WES was compared to reads at the same location in RNAseq. Based on the number of alternate and reference reads, a power calculation was obtained with beta-binomial distribution (power threshold used was greater than 80%). Mutation calls were deemed validated if 2 or greater alternate allele reads were observed in RNA-Seq at the site, as long as RNAseq was powered to detect an event at the specified location.

**FACS validation of ploidy estimates with ABSOLUTE:** Consistent with published studies of CLL(Brown et al., 2012; Edelmann et al., 2012), ABSOLUTE measured all CLL samples to be near diploid (**Table S3B**; median - 2, range 1.95-2.1). We confirmed the measurements using a standard assay for measuring DNA content. For this analysis, peripheral blood mononuclear cells from normal volunteers and CLL patients and cell lines are first stained with anti-CD5 FITC and anti-CD19 PE antibodies in a PBS buffer containing 1% BSA for 30 minutes on ice. After extensive washes, the cells were then stained with a PBS buffer contained 1% BSA, 0.03% saponin (Sigma) and 250ug/ml 7-AAD (Invitrogen) for 1 hour on ice, followed by analysis on a Beckman Coulter FC500 machine (**Figure S2A**).

**Estimation of mutation cancer cell fraction using ABSOLUTE:** We used the ABSOLUTE algorithm to calculate the purity, ploidy, and absolute DNA copy-numbers of each sample (Carter et al., 2012). Modifications were made to the algorithm, which are implemented in version 1.05 of the software, available for download at https://confluence.broadinstitute.org/display/CGATools/ABSOLUTE. Specifically, we added to the ability to determine sample purity from sSNVs alone, in samples where no sCNAs are present (the ploidy of such samples is 2*N*). In addition, estimates of sample purity and absolute copy-numbers are used to compute distributions over cancer cell fraction (CCF) values of each sSNV, as described (**Experimental Procedures**), and for sCNAs (described below).

The current implementation of ABSOLUTE does not automatically correct for sCNA subclonality when computing CCF distributions of sSNVs (this is an area of ongoing development). Fortunately, the few sCNAs that occurred in our CLL samples were predominantly clonal. Manual corrections were made for CLL driver sSNVs occurring at site of subclonal sCNAs (5 *TP53* sSNVs and 1 *ATM* sSNV), based on the sample purity, allelic fraction and the copy ratio of the matching sCNA.

Each sSNV was classified as clonal or subclonal based on the probability that the CCF exceeded 0.95. A probability threshold of 0.5 was used throughout the manuscript. However, as the histogram in **Figure 2A** shows, the distribution of events around the threshold was observed to be fairly uniform and results were not significantly affected across a range of thresholds. For example, the results of our analyses were unchanged when we altered our definition of clonal mutations to be (Pr(CCF>0.95)) > 0.75, and subclonal when Pr(CCF>0.95) was < 0.25, leaving uncertain mutations unclassified. Using these thresholds, CLLs with mutated *IGHV* and age were associated with a higher number of clonal mutations (*P* values of 0.05 and <0.0001, respectively). CLLs treated prior to sample collection had a higher number of subclonal mutations (*P*=0.01) and the subclonal set was enriched with putative drivers (*P* =0.0019). Importantly, the results of the clinical analysis also remained unchanged. FFS_Rx was shorter in samples in which a subclonal driver was detected (*P*=0.007) and regression models examining known poor prognostic indicators in CLL yielded an adjusted *P* value of 0.009.

One of the recurrent CLL cancer genes, *NOTCH1*, had 15 mutations, 14 of which were the identical canonical 2 base-pair deletions. Unlike sSNVs, the observed allelic fractions of indels events were not modeled as binomial sampling of reference and alternate sequence reads according to their true concentration in the sample(Carter et al., 2012). This was due to biases affecting the alignment of the short sequencing reads, which generally favor reference over alternate alleles. To measure the magnitude of this effect, we examined the allelic fraction (AF) of 514 germline 2bp deletions called in 4 normal germline WES samples. We observed that the distribution (data not shown) of allelic-fractions for heterozygous events was peaked at 0.41, as opposed to the expected mode of 0.5, with nearly all AFs between 0.3 to 0.6. Therefore, the bias factor towards reference is peaked at 0.82 but may range from 0.6 to 1 (unlikely to be greater than 1). CCF distributions for the 14 somatic indels in *NOTCH1* were calculated using bias

factors of 1.0 (no bias), 0.82 (bias point-estimate), and 0.6 (worst case observed). Reassuringly, the classification of *NOTCH1* indels as clonal or subclonal was highly robust and was essentially the same using the three values -- only a single case (CLL155) was ambiguous and was classified as subclonal using 1.0 and 0.82, and clonal using 0.6. Taking a conservative approach, not classifying a mutation as sub-clonal unless there is clear evidence for it, we decided to call this event as clonal for downstream analysis.

Estimation of CCF values for subclonal sCNAs is implemented (ABSOLUTEv1.05) in a manner analogous to the procedure for sSNVs (**Experimental Procedures**), although the transformation is more complex, due to the need for assumptions of the subclonal structure and the error model of microarray based copy-number data. Segmental sCNAs are defined as subclonal based on the mixture model used in ABSOLUTE (Carter et al., 2012). Let the functions $h(x)$ and $h'(x)$ denote a variance stabilizing transformation and its derivative, respectively. For SNP microarray data, these are defined as: $h(x) = \sinh^{-1}(bx)$, where $b = \frac{\left(e^{\sigma_\eta^2}-1\right)^{\frac{1}{2}}}{\sigma_\epsilon}$, and $h'(x) = \frac{b}{(1+(bx)^2)^{\frac{1}{2}}}$ (Huber et al., 2002). The values $\sigma_\epsilon$ and $\sigma_\eta$ denote additive and multiplicative noise scales, respectively, for the microarray hybridization being analyzed; these are estimated by HAPSEG (Carter et al., 2011). The calibrated probe-level microarray data become approximately normal under this transformation, which is used by HAPSEG to estimate the segmental allelic copy-ratios $r_i$ and the posterior standard deviation of their mean (under the transformation), $\sigma_i$ (Carter, 2011). An additional parameter $\sigma_H$ is estimated by ABSOLUTE(Carter et al., 2012), which represents additional sample-level variance corresponding to regional biases not captured in the probe-level model.

For a subclonal segment *i*, let $q_c$ denote the absolute copy number in the unaffected cells, and $q_s$ denote the absolute copy number in the altered cells. Both of these values are unknown but we used a simplifying assumption that the difference between $q_c$ and $q_s$ is one copy with $q_c$ being closer to the modal copy-number. Therefore, for subclonal deletions (copy ratios below the ratio of modal copy number), $q_s$ was set to the nearest copy number below the measured value, and $q_c$=$q_s$+1. For subclonal gains (ratios above the modal number), $q_s$ was set to the nearest copy number above the measured value, and $q_c$=$q_s$-1. Because the CLL genomes analyzed here were universally near diploid, this was nearly equivalent to assuming that subclonal deletions had $q_s$=0 in the affected cells and gains $q_s$=2, with $q_c$=1 in both cases (in allelic units). However, we note that these assumptions would not be strictly correct in genomes after doubling, or in cases of high-level amplification. In these cases, calculation of posterior CCF distributions will require integration over $q_s$ and $q_c$, averaging over the set of plausible subclonal genomic configurations.

Let $r_c$ and $r_s$ be the theoretical copy ratio values corresponding to $q_c$ and $q_s$ (accounting for sample purity, ploidy, and the modeled attenuation rate of the microarray (Carter et al., 2011; Carter et al., 2012)). Let $d = r_s - r_c$, then, for CCF $c$, let $r_x(c) = dc + r_c$. Then $P(c) \propto \mathcal{N}\left(h\left(r_x(c)\right) \mid h(r_i), (\sigma_i + \sigma_H)^2\right) \left|h'\left(r_x(c)\right)\right|$. The absolute value of the derivative is required due to the change of coordinates from *x* to *h(x)*. The distribution over CCF is

obtained by calculating these values over a regular grid of 100 $c$ values and normalizing. We note that, when copy numbers are estimated directly from sequencing data, the calculation is simpler, as there is no attenuation effect and $h(x) = x$. These calculations were used to generate the 95% confidence intervals on the CCF of subclonal driver sCNAs shown in **Figure 4** and **Figure S4A,B**.

**Cancer gene census list and conservation annotations:** Conservation of a specific mutated site was adapted from UCSC conservation score track. A scale of 0-100 was linearly converted from the -6 to 6 scale used in the phastCons track(Siepel et al., 2005). To confirm that driver mutations are more likely to occur in conserved sites, we quantified the conservation in the COSMIC database (Forbes et al., 2008) hotspots and compared it to non-COSMIC hotspots coding location. We matched conservation information for 5085 sites that had greater than 3 exact hits reported in mutations deposited in the COSMIC database, and compared it to conservation found for a set of non-overlapping 5085 randomly sampled coding sites. The conservation was higher in the COSMIC sites than in the non-COSMIC coding sites set (mean conservation 82.39 and 62.15, respectively, p<1e-50). We noted that the distribution of events was not uniform, and nearly one half of COSMIC hotspots had a conservation measure greater than 95 (49.65%, compared to 15.5% in the non-COSMIC set, p<1e-50). For our calculations, we used a cut off of >95 to designate conserved sites likely to contain higher proportion of cancer drivers. We complemented the analysis for putative driver event enrichment by matching the altered genes to the Cancer Gene Census (Futreal et al., 2004).

**Clustering analysis of sSNVs in 18 CLL sample pairs:** In order to better resolve the true cancer cell fraction (CCF) of sSNVs detected in longitudinal samples, we employed a previously described Bayesian clustering procedure (Escobar and West, 1995). This approach exploits the assumption that the observed subclonal sSNV CCF values were sampled from a smaller number of subclonal cell populations (subclones). All remaining uncertainty (including the exact number of clusters) was integrated out using a mixture of Dirichlet processes, which was fit using a Gibbs sampling approach, building on a previously described framework (Escobar and West, 1995).

The inputs to this procedure are the posterior CCF distributions for each sSNV being considered. We note that the CCF distributions for sCNAs could be added into the model, however we did not attempt this in the present study. CCF distributions are represented as 100-bin histograms over the unit interval; the two-dimensional CCF distributions used for the 2D clustering of longitudinal samples were obtained as the outer product of the matched histogram pairs for each mutation, resulting in 10,000-bin histograms (**Figure S4**). We note that the use of histograms to represent posterior distributions on CCF, although computationally less efficient than parametric forms, has the advantage that CCFs of different mutation classes may be easily combined in the model, even though their posteriors may have very different forms. We also note that the algorithm implementation is identical for the single sample and paired (longitudinal) sample cases, although only the latter was used in the present study.

At each iteration of the Gibbs sampler, each mutation is assigned to a unique cluster and the posterior CCF distribution of each cluster is computed using Bayes' rule, as opposed to drawing a sample from the posterior (a uniform prior on CCF from 0.01 to 1 is used). When considering the probability of a mutation to join an existing cluster, the likelihood calculation of the mutation arising from the cluster is integrated over the uncertainty in the cluster CCF. This allows for rapid convergence of the Gibbs sampler to its stationary distribution, which was typically obtained in fewer than 100 iterations for the analysis presented in this study. We ran the Gibbs sampler for 1,000 iterations, of which the first 500 were discarded before summarization.

Because of the small number of clonal mutations in some WES samples, we make an additional modification to the standard Dirichlet process model by adding a fixed clonal cluster that persists even if no mutation is assigned to it. This reflects our prior knowledge that clonal mutations must exist, even if they are the minority of detected mutations. For the samples analyzed here, this modification had very little effect.

A key aspect of implementing the Dirichlet process model on WES datasets is re-parameterization of prior distributions on the number of subclones $k$ as priors on the concentration parameter $\alpha$ of the Dirichlet process model. Importantly, this must take into account the number of mutations $N$ input to the model, as the effect of $\alpha$ on $k$ is strongly dependent on $N$ (Escobar and West, 1995). We accomplish this by constructing a map from a regular grid over $\alpha$ to expected values of $k$, given $N$, using the fact that: $P(k|\alpha, N) = c_N(k)N!\,\alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha+N)}$ (Escobar and West, 1995), where the $c_N(k)$ factors correspond to the unsigned Stirling numbers of the first kind. With this map in hand, we perform an optimization procedure to find parameters $a$ and $b$ of a prior Gamma distribution over $\alpha$ resulting in the minimal Kullback-Leibler divergence with the specified prior over $k$ (the divergence was computed numerically on the histograms). Once the prior over $\alpha$ has been represented as a Gamma distribution, learning about $\alpha$ (and therefore $k$) from the data can be directly incorporated into the Gibbs sampling procedure, resulting in a continuous mixture of Dirichlet processes (Escobar and West, 1995). This allows consistent parameterization of prior knowledge (or lack thereof) on the number of subclonal populations in the face of vastly different numbers of input mutations, which is necessary for making consistent inferences across differing datasets (e.g. WES vs. WGS). We note that taking uncertainty about $\alpha$ into account is necessary for inferences on the number of subclonal populations to be strictly valid, since implementations with fixed values of $\alpha$ result in an implicit prior over $k$ that depends upon $N$ (this is especially important for smaller values of $N$). For the application presented in this study (**Figure 4**), we specified a weak prior on $k$ using a negative binomial distribution with $r$=10, $\mu$=2 (these values favored 1-10 clusters). We note that these are the only two parameters of the clustering analysis.

Upon termination of the Gibbs sampler, we summarized the posterior probability over the CCF of each sSNV by averaging the posterior cluster distribution for all clusters to which the sSNV was assigned during sampling. This allowed shrinkage of the CCF probability distributions (as shown in **Fig. 4;** pre-clustering results are shown in **Fig. S4A-B**), without having to choose an exact number of subclonal clusters. Note that the 18 longitudinal sample pairs contain 1 CLL sample pair not initially included in the 160

CLLs (CLL020).

**Gene Expression Profiling**: Total RNA was isolated from viably frozen PBMCs or B cells from CLL patients that were followed longitudinally (Midi kit; Qiagen, Valencia CA), and hybridized to the U133Plus 2.0 array (Affymetrix, Santa Cruz, CA) at the DFCI Microarray Core Facility. All expression profiles were processed using RMA, implemented by the PreprocessDataset module in GenePattern (http://www.broadinstitute.org/cancer/software/genepattern/) (Irizarry et al., 2003; Reich et al., 2006). Probes were collapsed to unique genes by selecting the probe with the maximal average expression for each gene. Batch effects were further removed using the ComBat module in GenePattern(Johnson et al., 2007) (Reich et al., 2006). Visualizations in GENE-E (http://www.broadinstitute.org/cancer/software/GENE-E) were based on logarithmic transformation (log2) of the data and centering each gene (zero mean). These data can be accessed at http://www.ncbi.nlm.nih.gov/geo/info/linking.html with accession number GSE37168.

**RNA pyrosequencing for mutation confirmation**: Quantitative targeted sequencing to detect somatic mutation within cDNA was performed, as previously described (Armistead et al., 2008). In brief, biotinylated amplicons generated from PCR of the regions of transcript surrounding the mutation of interest were generated. Immobilized biotinylated single-stranded DNA fragments were isolated per manufacturer's protocol, and sequencing undertaken using an automated pyrosequencing instrument (PSQ96; Qiagen, Valencia CA), followed by quantitative analysis using Pyrosequencing software (Qiagen).

**Statistical methods** Statistical analysis was performed with MATLAB (MathWorks, Natick, MA), R version 2.11.1 and SAS version 9.2 (SAS Institute, Cary, NC). Categorical variables were compared using the Fisher Exact test, and continuous variables were compared using the Student's t-test, Wilcoxon rank sum test, or Kruskal Wallis test as appropriate; the association between two continuous variables was assessed by the Pearson correlation coefficient. The time from the date of sample to first therapy or death (failure-free survival from sample time or FFS_Sample) was calculated as the time from sample to the time of the first treatment after the sample or death and was censored at the date of last contact. FFS_Rx (failure-free survival from first treatment after sampling) was defined as the time to the $2^{nd}$ treatment or death from the $1^{st}$ treatment following sampling, was calculated only for those patients who had a $1^{st}$ treatment after the sample and was censored at the date of last contact for those who had only one treatment after the sample. Time to event data were estimated by the method of Kaplan and Meier, and differences between groups were assessed using the log-rank test. Unadjusted and adjusted Cox modeling was performed to assess the impact of the presence of a subclonal driver and a driver irrespective of the CCF on FFS_Sample and FFS_Rx. A chi-square test with 1 degree of freedom and the -2 Log-likelihood statistic was used to test the prognostic independence of subclonal status in Cox modeling using a full model and one without subclonal status included. We also formally tested for non-proportionality of the hazards in **Figure 6B** First, we plotted the log(-log(survival) versus

log(time) for the two categories, and demonstrated that curves do not cross, which supports the fact that they are proportional. Second, we also tested for non-proportionality by including a time varying covariate for each variable in the model. None of these were significant indicating that the hazards are proportional. Models were adjusted for known prognostic factors for CLL treatment including the presence of a 17p deletion, the presence of a 11q deletion, *IGHV* mutational status, and prior treatment at the time of sample. Cytogenetic abnormalities were primarily assessed by FISH and if unknown, genomic data were included. For unknown *IGHV* mutational status an indicator was included in adjusted modeling and was not found to be significant. All P-values are two-sided and considered significant at the 0.05 level unless otherwise noted.

## D. References


Armistead, P., Mohseni, M., Gerwin, R., Walsh, E., Iravani, M., Chahardouli, B., Rostami, S., Zhang, W., Neuberg, D., Rioux, J.*, et al.* (2008). Erythroid-lineage-specific engraftment in patients with severe hemoglobinopathy following allogeneic hematopoietic stem cell transplantation. Exp Hematol. *36*, 1205-1215.

Berger, M.F., Lawrence, M.S., Demichelis, F., Drier, Y., Cibulskis, K., Sivachenko, A.Y., Sboner, A., Esgueva, R., Pflueger, D., Sougnez, C.*, et al.* (2011). The genomic complexity of primary human prostate cancer. Nature. *470*, 214-220.

Brown, J.R., Hanna, M., Tesar, B., Werner, L., Pochet, N., Asara, J.M., Wang, Y.E., Dal Cin, P., Fernandes, S.M., Thompson, C.*, et al.* (2012). Integrative genomic analysis implicates gain of PIK3CA at 3q26 and MYC at 8q24 in chronic lymphocytic leukemia. Clin Cancer Res. *18*, 3791-3802.

Carter, S.L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P.W., Onofrio, R.C., Winckler, W., Weir, B.A.*, et al.* (2012). Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol. *30*, 413-421.

Carter, S.L., Meyerson, M., & Getz, G. (2011). Accurate estimation of homologue-specific DNA concentration ratios in cancer samples allows long-range haplotyping. . Available from Nature Precedings <http://hdlhandlenet/10101/npre201164941%3E.

Chapman, M.A., Lawrence, M.S., Keats, J.J., Cibulskis, K., Sougnez, C., Schinzel, A.C., Harview, C.L., Brunet, J.P., Ahmann, G.J., Adli, M.*, et al.* (2011). Initial genome sequencing and analysis of multiple myeloma. Nature. *471*, 467-472.

Cibulskis, K., McKenna, A., Fennell, T., Banks, E., DePristo, M., and Getz, G. (2011). ContEst: estimating cross-contamination of human samples in next-generation sequencing data. Bioinformatics. *27*, 2601-2602.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M.*, et al.* (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. *43*, 491-498.

Edelmann, J., Holzmann, K., Miller, F., Winkler, D., Buhler, A., Zenz, T., Bullinger, L., Kuhn, M.W., Gerhardinger, A., Bloehdorn, J.*, et al.* (2012). High-resolution genomic profiling of chronic lymphocytic leukemia reveals new recurrent genomic alterations. Blood. [Epub ahead of print].

Escobar, M., and West, M. (1995). Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association. *90*, 577-588.

Eskandarpour, M., Huang, F., Reeves, K.A., Clark, E., and Hansson, J. (2009). Oncogenic NRAS has multiple effects on the malignant phenotype of human melanoma cells cultured in vitro. Int J Cancer. *124*, 16-26.

Fisher, R.A. (1932). Statistical methods for research workers, 4th edn (Oliver and Boyd).

Fisher, S., Barry, A., Abreu, J., Minie, B., Nolan, J., Delorey, T.M., Young, G., Fennell, T.J., Allen, A., Ambrogio, L.*, et al.* (2011). A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. Genome Biol. *12*, R1.

Forbes, S.A., Bhamra, G., Bamford, S., Dawson, E., Kok, C., Clements, J., Menzies, A., Teague, J.W., Futreal, P.A., and Stratton, M.R. (2008). The Catalogue of Somatic Mutations in Cancer (COSMIC). Curr Protoc Hum Genet. *Chapter 10*, Unit 10 11.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004). A census of human cancer genes. Nat Rev Cancer. *4*, 177-183.

Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics. *18 Suppl 1*, S96-104.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. *4*, 249-264.

Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. *8*, 118-127.

Lohr, J.G., Stojanov, P., Lawrence, M.S., Auclair, D., Chapuy, B., Sougnez, C., Cruz-Gordillo, P., Knoechel, B., Asmann, Y.W., Slager, S.L.*, et al.* (2012). Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. Proc Natl Acad Sci U S A. *109*, 3879-3884.

Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. *12*, R41.

Rassenti, L., Jain, S., Keating, M., Wierda, W., Grever, M., Byrd, J., Kay, N., Brown, J., Gribben, J., Neuberg, D.*, et al.* (2008). Relative value of ZAP-70, CD38, and immunoglobulin mutation status in predicting aggressive disease in chronic lymphocytic leukemia. Blood. *112*, 1923-1930.

Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P., and Mesirov, J.P. (2006). GenePattern 2.0. Nat Genet. *38*, 500-501.

Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat Biotechnol. *29*, 24-26.

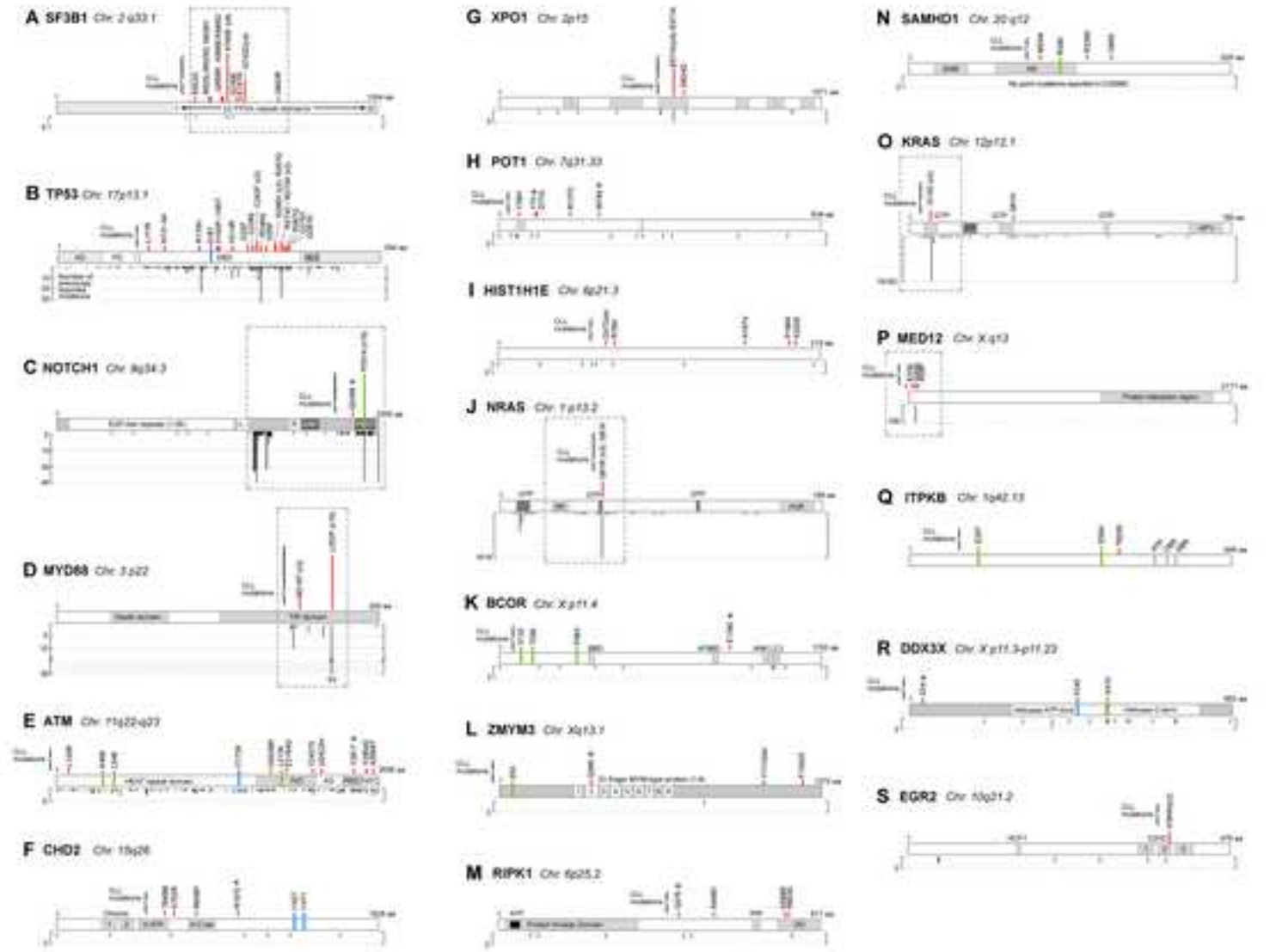
Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S.*, et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*, 1034-1050.

Smoley, S.A., Van Dyke, D.L., Kay, N.E., Heerema, N.A., Dell' Aquila, M.L., Dal Cin, P., Koduru, P., Aviram, A., Rassenti, L., Byrd, J.C.*, et al.* (2010). Standardization of fluorescence in situ hybridization studies on chronic lymphocytic leukemia (CLL) blood and marrow cells by the CLL Research Consortium. Cancer Genet Cytogenet. *203*, 141-148.

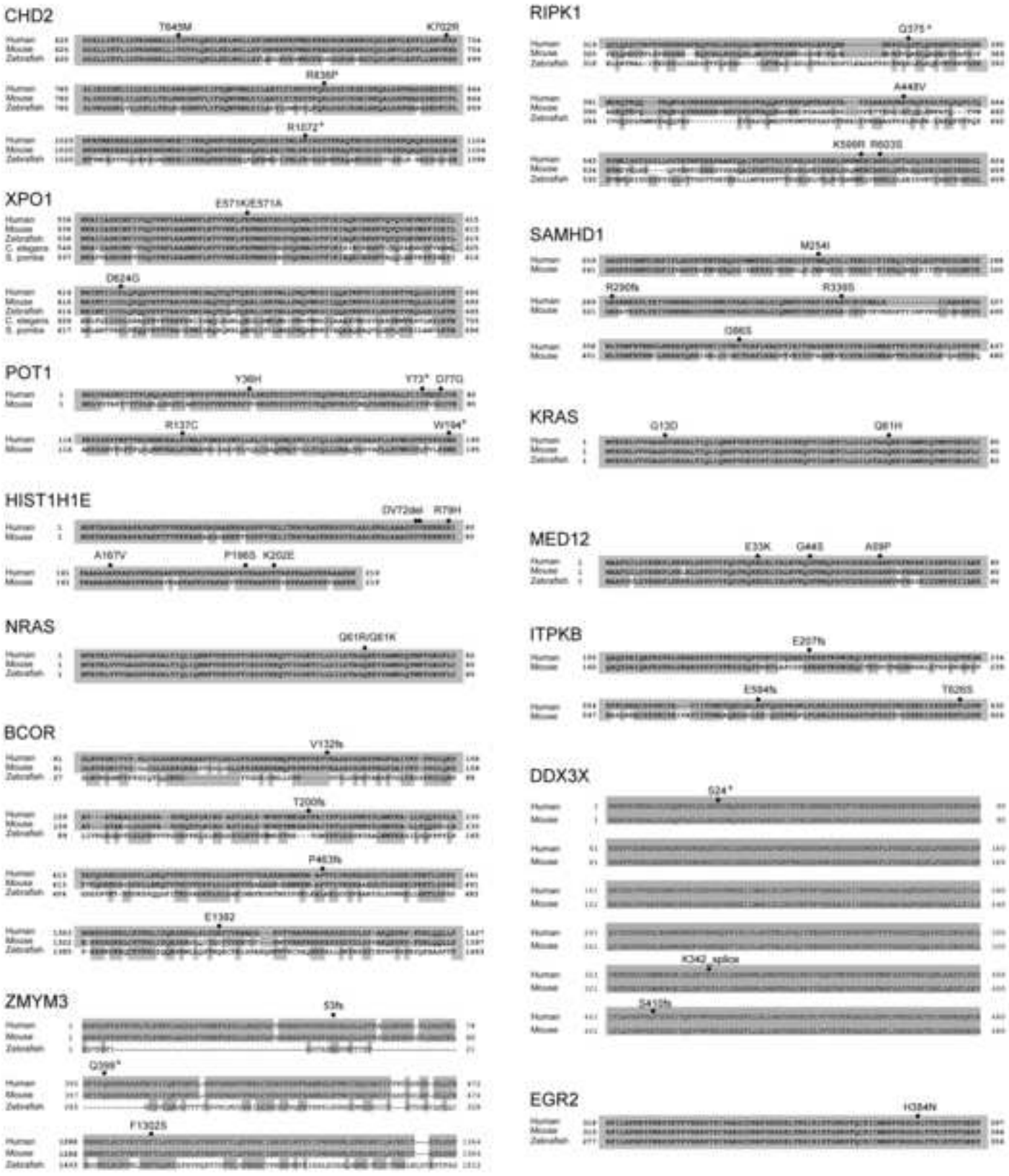
Wang, L., Lawrence, M.S., Wan, Y., Stojanov, P., Sougnez, C., Stevenson, K., Werner, L., Sivachenko, A., DeLuca, D.S., Zhang, L.*, et al.* (2011). SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. N Engl J Med. *365*, 2497-2506.

Yoshida, K., Sanada, M., Shiraishi, Y., Nowak, D., Nagata, Y., Yamamoto, R., Sato, Y., Sato-Otsubo, A., Kon, A., Nagasaki, M.*, et al.* (2011). Frequent pathway mutations of splicing machinery in myelodysplasia. Nature. *478*, 64-69.

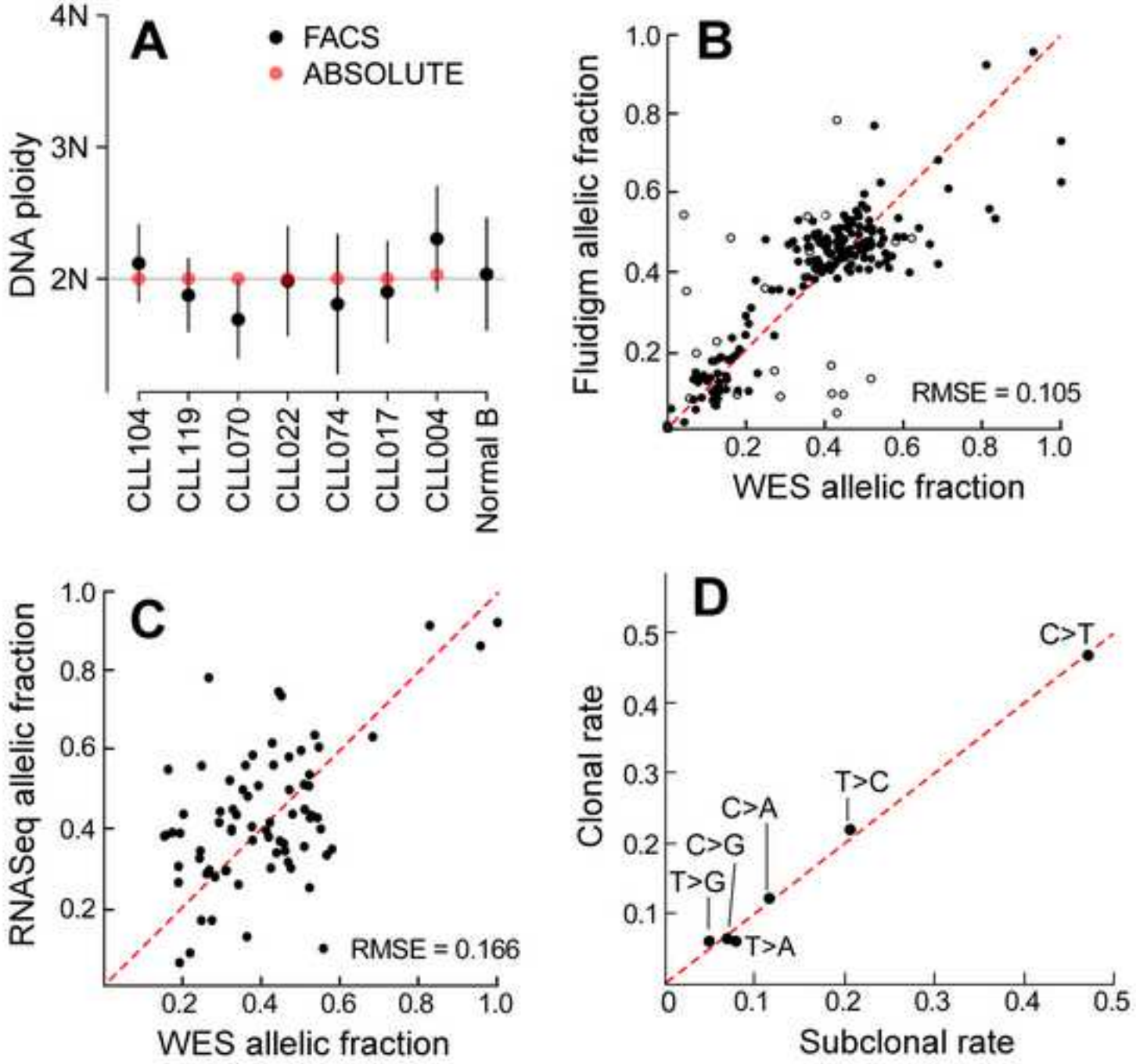
**Supplemental Figure**



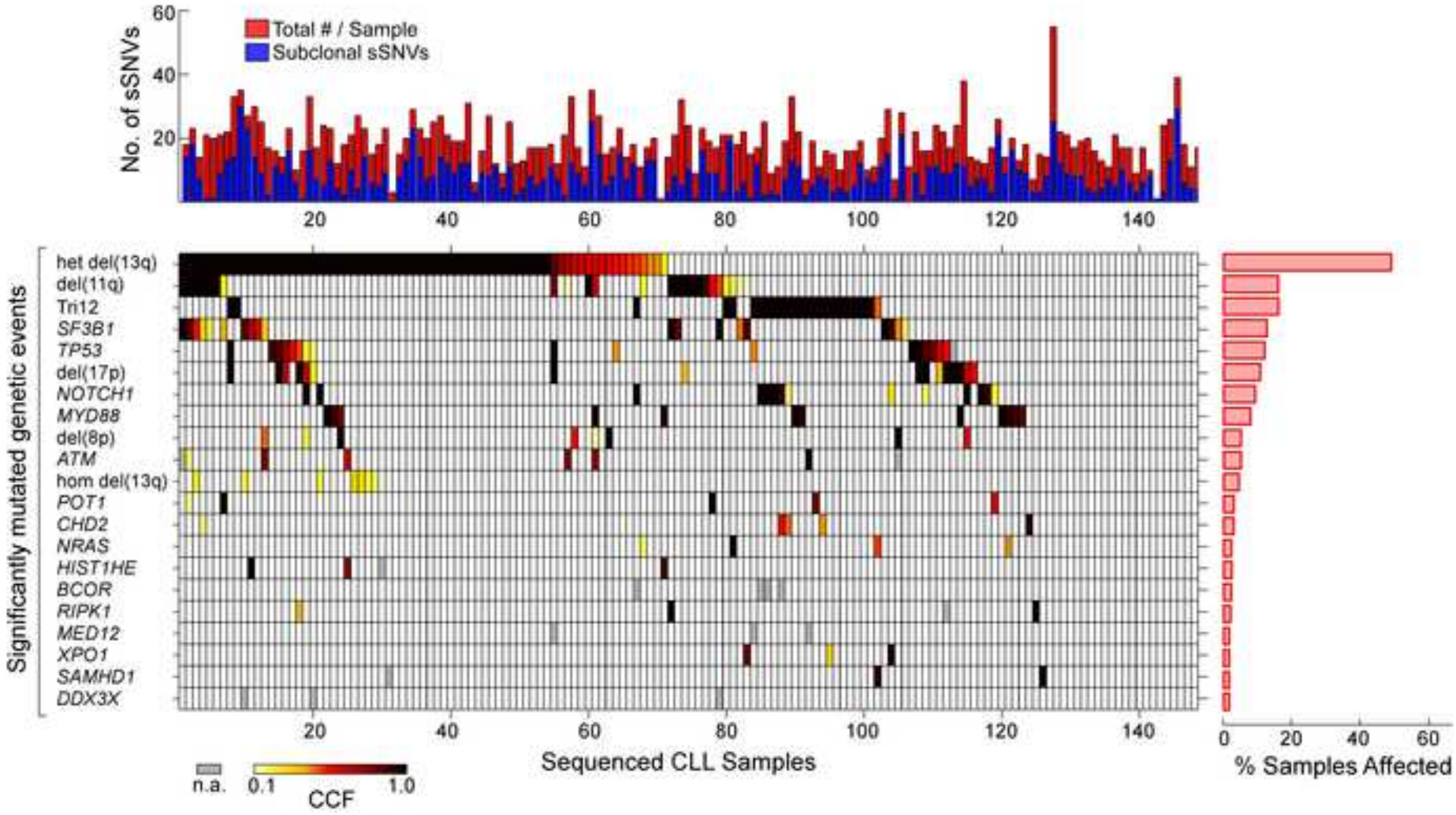
**Supplemental Figure**



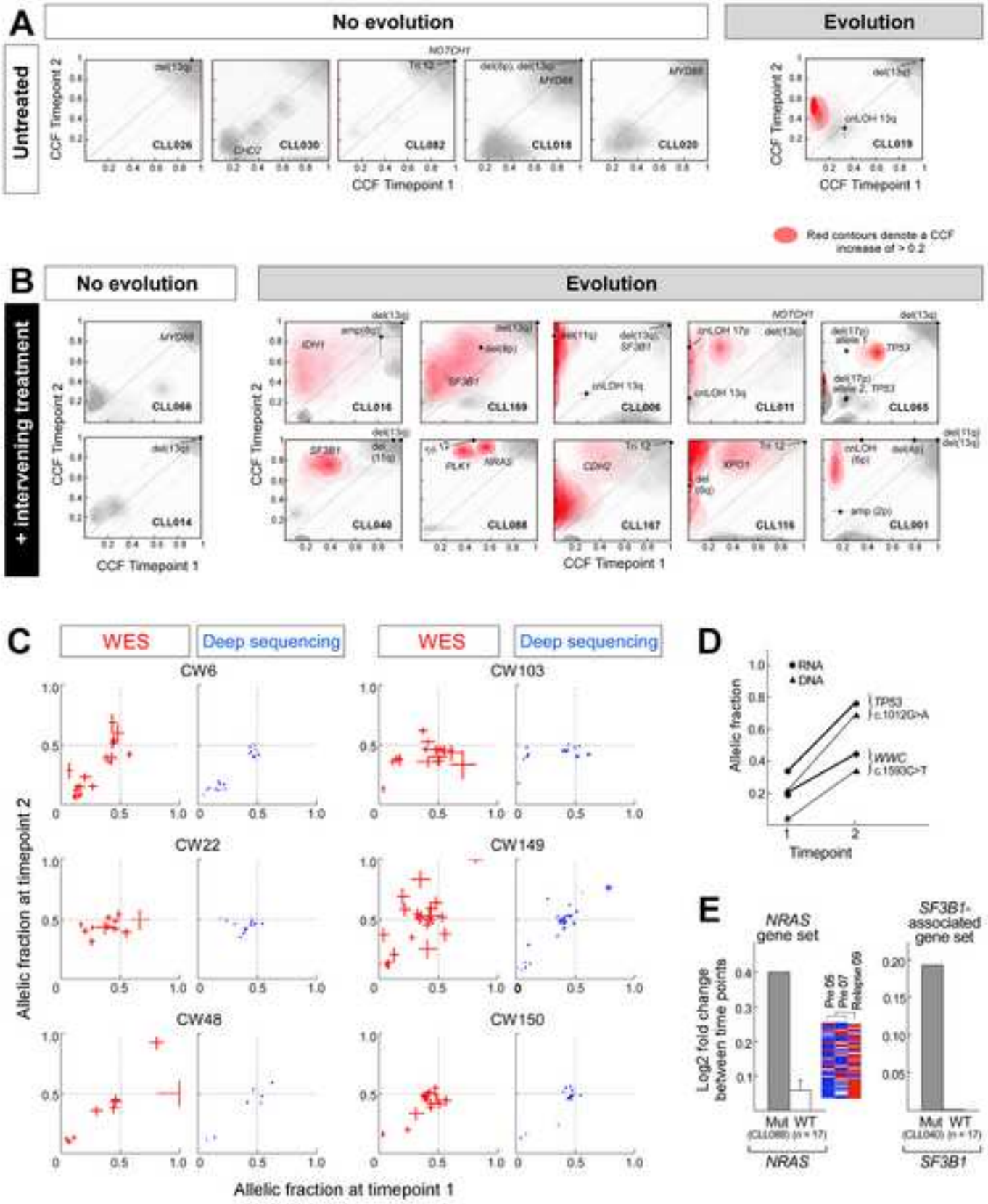
**Supplemental Figure**



**Supplemental Figure**



**Supplemental Figure**



**Supplemental Figure**